AD-A100 606 WISCONSIN UNIV-MADISON MATHEMATICS RESEARCH CENTER F/G 12/1
ANDREWS' PLOTS AND THEIR APPLICATIONS.(U)
APR 81 A M HERZBERG DAAG29-80-C-0041
UNCLASSIFIED MRC-TSR-2203 NL

1 OF 1
AD
A100606

END
DATE
FILMED
7-81
DTIC

MRC Technical Summary Report #2203

ANDREWS' PLOTS AND THEIR APPLICATIONS

Agnes M. Herzberg

Mathematics Research Center
University of Wisconsin—Madison
610 Walnut Street
Madison, Wisconsin 53706

April 1981

Received March 9, 1981

Approved for public release
Distribution unlimited

Sponsored by

U. S. Army Research Office
P. O. Box 12211
Research Triangle Park
North Carolina 27709

2

UNIVERSITY OF WISCONSIN - MADISON
MATHEMATICS RESEARCH CENTER

ANDREWS' PLOTS AND THEIR APPLICATIONS

Agnes M. Herzberg[†]

Technical Summary Report #2203
April 1981

ABSTRACT

A brief survey of several graphical multivariate techniques are given. Andrews' method is exploited as a graphical tool for the examination of changes over time in the parameters of a time series model. An example is given to illustrate the method

AMS(MOS) Subject Classification: 62H30, 62M10

Key Words: Andrews' plots, trees, castles, Chernoff's faces, ideographs, time series, cyclic, trend, exploratory analysis.

Work Unit No. 4 - Statistics and Probability

[†]Imperial College of Science and Technology, University of London, London, England.

Sponsored by the United States Army under Contract No. DAAG29-80-C-0041.

SIGNIFICANCE AND EXPLANATION

A brief survey of several graphical multivariate techniques are given. One of these due to Andrews is given in more detail. In his method, Andrews represents each multidimensional point by a Fourier function. The clustering of plots of these functions is equivalent to the clustering of the multidimensional points. Andrews' method is exploited as a graphical tool for the examination of changes over time in the parameters of a time series model. An example consisting of temperature data is given to illustrate the method.

The responsibility for the wording and views expressed in this descriptive summary lies with MRC, and not with the author of this report.

ANDREWS' PLOTS AND THEIR APPLICATIONS

Agnes M. Herzberg[†]

1. Introduction.

If multivariate data are m-dimensional, then each set of m measurements can be represented as an m-dimensional point. For $m = 1,2$, these points may be plotted and clusters easily determined by inspection. For $m > 2$, this is more difficult. Several authors have developed graphical techniques to plot high-dimensional data in two dimensions in order to be able to visually cluster the data; see, for example, Andrews (1972), Chernoff (1973), Kleiner and Hartigan (1981) and Anderson (1928, 1936). More mathematical techniques have been given by Beale (1969) and Banfield and Bassil (1977).

2. Several graphical methods.

Kleiner and Hartigan (1981) introduced what they termed *trees* and *castles*. First, a hierarchical clustering algorithm is applied to the m variables over all the points; see for example Gnanadesikan (1977). From this the structure of the tree or castle will be determined. All points will be represented by a similar structure, i.e. the thickness, position and angle of the branches in the case of trees will be the same, but the length of the branches will be determined by the sizes of the respective variables for the individual points. Similar trees and castles determined by visual inspection are clustered.

[†]Imperial College of Science and Technology, University of London, London, England.

Sponsored by the United States Army under Contract No. DAAG29-80-C-0041.

Chernoff (1973) represents each variable as a feature of a *face* , for example length of mouth, shape of mouth, size of eyes, etc. The resulting clustering of this representation is very subjective because different people focus on different features of faces.

Anderson (1928, 1936) was very concerned with the sepal length and width and petal length and width of irises. He developed pictorial methods which he called *ideographs* for representing and comparing these four-dimensional data. An ideograph looks like an upside-down U with some width. In the case of the iris measurements, the inside and outside height and width measurements of the ideograph are proportional to the sepal length and width and the petal length and width, respectively. Similar ideographs can easily be clustered by visual inspection.

There are many other graphical representations for multivariate data, but these will not be discussed.

3. Andrews' plots.

Andrews (1972) proposed the following simple and useful method of plotting high-dimensional data in two dimensions. If the data are m-dimensional, each point $\underset{\sim}{x}' = (x_1,\cdots,x_m)$, where $x_i (i = 1,\cdots,m)$ are the measured variables, is represented by the function

$$f_{\underset{\sim}{x}}(t) = x_1 2^{-\frac{1}{2}} + x_2 \sin t + x_3 \cos t + x_4 \sin 2t + x_5 \cos 2t + \cdots \qquad (1)$$

plotted over the range $-\pi < t < \pi$. The functions given by (1) have several properties. If $\underset{\sim}{x}_i = (x_{1i},\cdots,x_{mi}) \quad (i = 1,\cdots,n)$ are n points in m-dimensional space, then

$$f_{\bar{\underset{\sim}{x}}}(t) = \frac{1}{n} \sum_{i=1}^{n} f_{\underset{\sim}{x}_i}(t)$$

$$\| f_{\underset{\sim}{x}_i}(t) - f_{\underset{\sim}{x}_j}(t) \|_{L_2} = \int_{-\pi}^{\pi} \{ f_{\underset{\sim}{x}_i}(t) - f_{\underset{\sim}{x}_j}(t) \}^2 dt$$

$$= \pi \| \underset{\sim}{x}_i - \underset{\sim}{x}_j \|^2$$

$$= \pi \sum_{k=1}^{m} (x_{ki} - x_{kj})^2 .$$

Thus Andrews' plots will preserve means, distances and variances and will also give one-dimensional projections. When (1) is plotted for each data point $\underset{\sim}{x}$, the clustering of the points may be seen by a banding together of the plots of the functions. Since the functions preserve the distance property, plots of the functions that are close together imply that the corresponding data points are close together.

4. Variation of model parameters.

Herzberg and Hickie (1981) considered the following. Let the regression model be written in the form

$$\underset{\sim}{Y}_j = \underset{\sim}{X}\, \underset{\sim}{\beta}_j + \underset{\sim}{U}_j \quad (j = 1, \cdots, T - n + 1), \tag{2}$$

where T is the total number of observations n is the number of observations in each subgroup of observations used for estimating the unknown parameters, $\underset{\sim}{Y}_j = (y_{1j}, \cdots, y_{nj})'$ is an $n \times 1$ vector, y_{kj} being the k^{th} observation in the j^{th} subgroup $(k = 1, \cdots, n)$, $\underset{\sim}{X}$ is the $n \times m$ matrix of the regressors, $\underset{\sim}{\beta}_j$ is the $m \times 1$ vector of unknown parameters and $\underset{\sim}{U}_j$ is the $n \times 1$ vector of error terms. All the elements of the $\underset{\sim}{U}_j$'s are assumed to be independent and normally distributed with mean 0 and variance σ^2 . It is assumed that the T observations are taken sequentially over time and it is desired to examine the variations in the $\underset{\sim}{\beta}_j$ over time.

Let $\hat{\beta}_j = (\hat{\beta}_{1j},\cdots,\hat{\beta}_{mj})'$ be the $m \times 1$ vector of least squares estimates of the elements of the vector β_j obtained from the j^{th} set of n observations $(n < T)$, i.e. $\hat{\beta}_1$ is estimated from the first n observations, $\hat{\beta}_2$ is estimated from the second observation to the $(n+1)^{st}$ observations, etc. From each $\hat{\beta}_j$ plot the function $f_{\hat{\beta}_j}(t)$, defined in (1), over the range $-\pi < t < \pi$. The plots of these functions will show the change over time in the vector of coefficients $\hat{\beta}_j$.

Herzberg and Hickie (1981) consider two sets of data using polynomial and Fourier series models in (2). One of the sets of data has a cyclic effect, the other having cyclic effect plus trend. For both sets of data it was known that the period was 12 months. It could also be seen that every 12^{th} plot was similar.

For one of their examples, namely the monthly mean daily air temperatures (°C) at sea level for England and Wales from January 1970 to December 1977 as published by the Central Statistical Office Monthly Digest of Statistics (HMSO), Herzberg and Hickie (1981) fitted the cubic polynomial model,

$$E(y_{j+i-1}) = \beta_{1j} + \beta_{2j}i + \beta_{3j}i^2 + \beta_{4j}\, i^3, \tag{3}$$

by least squares. Here y_{j+i-1} is the observed temperature in mouth $j + i-1$ For each j in (3) fixed, $i = 1,\cdots,12$, $\hat{\beta}_j = (\hat{\beta}_{1j},\hat{\beta}_{2j},\hat{\beta}_{3j},\hat{\beta}_{4j}\}'$ the least squares estimate of β_j was obtained $(j = 1,\cdots,85)$. Figure 1 shows the resulting Andrews' plots. The plots in Figure 1.k are those obtained from $\hat{\beta}_j$ $(j = k, k+12, k+24, k+48, k+60, k+72, k+84;\ j \leq 85)$. It can be seen that the plots in each of the Figures 1.k are similar.

In situations where the period is unknown, Andrews' plots may be plotted for several values of n in order to determine similarities and, therefore, the length of the period.

Because of their mathematical and resulting statistical properties, Andrews' plots can be used as a tool for finding outliers in a time series. Work is at present being done on this and on using Andrews' plots as a sequential graphical method for discriminating among models.

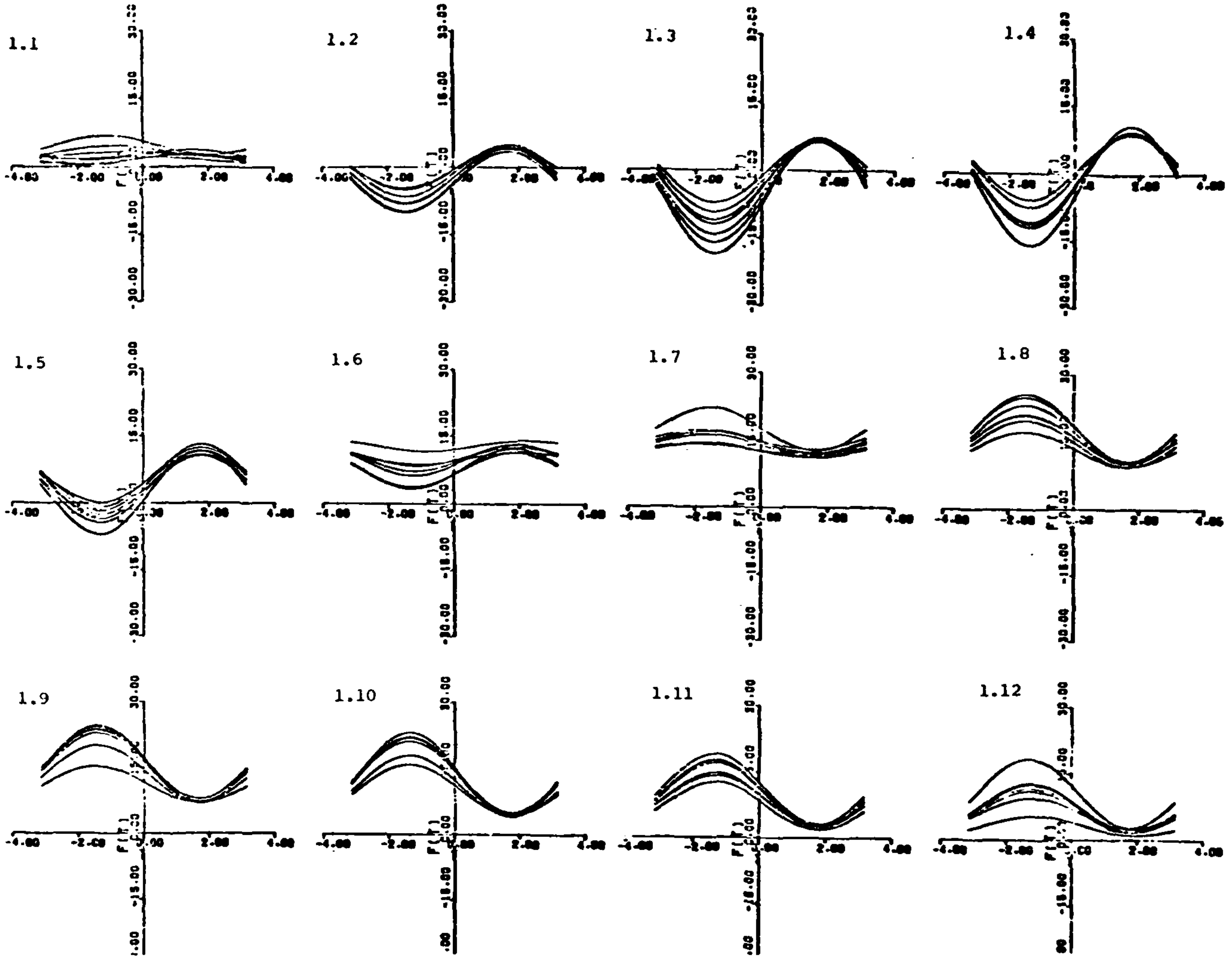

Figure 1. Andrews' plots $f_{\hat{\beta}_j}(t)$ $(j = 1,\cdots,85)$, $\hat{\underset{\sim}{\beta}}_j$ obtained from (3) using the monthly mean daily air temperature (°C) at sea level for England and Wales, January 1970 to December 1977; Figure 1.k $(k = 1,\cdots,12)$ consists of Andrews' plots for $j = k, k + 12, k + 24, k + 36, k + 48, k + 60, k + 72, k + 84$ $(j \leq 85)$.

REFERENCES

Anderson, E. (1928). The problem of species in the northern blue flags, *Iris versicolor L. and Iris virginica L*. Ann. Mo. Bot. Gard. **15**, 241-332.

Anderson, E. (1936). The species problem in *Iris*. Ann. Mo. Bot. Gard. **23**, 457-509.

Andrews, D. F. (1972). Plots of high-dimensional data. Biometrics **28**, 125-136.

Banfield, C. A. and Bassil, L. C. (1977). A transfer algorithm for non-hierarchical classification. Appl. Statist. **26**, 206-210.

Beale, E. M. L. (1969). Euclidean cluster analysis. Bull. Int. Statist. Inst. **43**, II, 92-94.

Chernoff, H. (1973). The use of faces to represent points in k-dimensional space graphically. J. Amer. Statist. Assoc. **68**, 361-368.

Gnanadesikan, R. (1977). *Methods for Statistical Data Analysis of Multivariate Observations*. New York: Wiley.

Herzberg, A. M. and Hickie, J. S. (1981). An investigation of Andrews' plots to show time variations of model parameters. Submitted.

Kleiner, B. And Hartigan, J. A. (1981). Representing points in many dimensions by trees and castles. J. Amer. Statist. Assoc. To appear.

AMH/db

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER 2203	2. GOVT ACCESSION NO. AD-A100 606	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) ANDREWS' PLOTS AND THEIR APPLICATIONS		5. TYPE OF REPORT & PERIOD COVERED Summary Report - no specific reporting period
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Agnes M. Herzberg		8. CONTRACT OR GRANT NUMBER(s) DAAG29-80-C-0041
9. PERFORMING ORGANIZATION NAME AND ADDRESS Mathematics Research Center, University of Wisconsin 610 Walnut Street Madison, Wisconsin 53706		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS #4 - Statistics and Probability
11. CONTROLLING OFFICE NAME AND ADDRESS U. S. Army Research Office P.O. Box 12211 Research Triangle Park, North Carolina 27709		12. REPORT DATE April 1981
		13. NUMBER OF PAGES 7
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) UNCLASSIFIED
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Andrews' plots, trees, castles, Chernoff's faces, ideographs, time series, cyclic, trend, exploratory analysis.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

A brief survey of several graphical multivariate techniques are given. Andrews' method is exploited as a graphical tool for the examination of changes over time in the parameters of a time series model. An example is given to illustrate the method.

DATE
ILMED

8